SIMULATION OF DAILY ACTIVITITIES AT REGIONAL CENTERS

MONARC Collaboration

Alexander Nazarenko and Krzysztof Sliwa

MONARC SIMULATION TOOLS

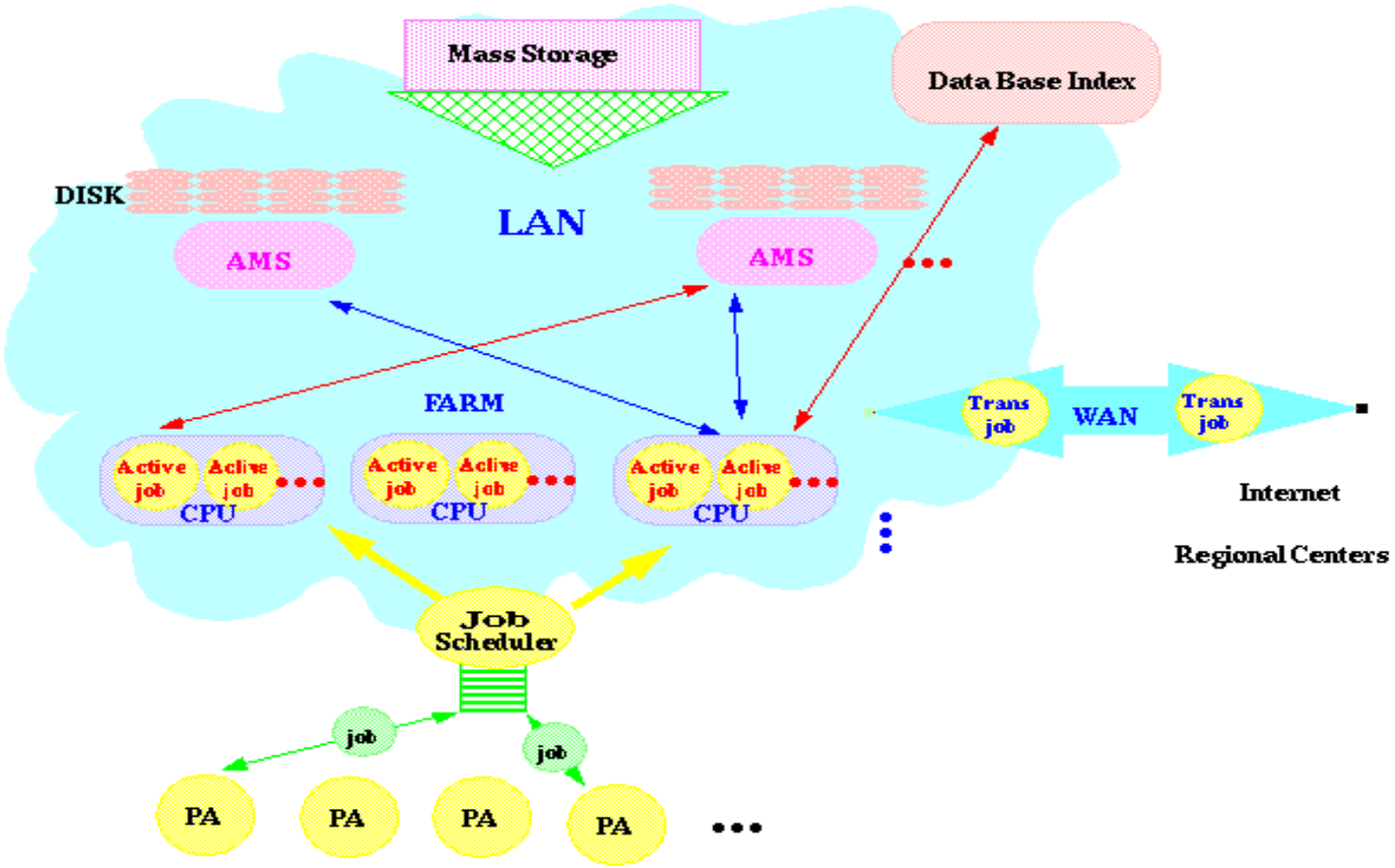
- **The simulation tool developed within MONARC is based on Java technology which provides adequate tools for developing a flexible and distributed process oriented simulation. Java has a built-in support for multi-threaded objects and for concurrent processing, which can be used for simulation purposes provided a dedicated scheduling mechanism is developed (this "simulation engine" has been developed by Iosif Legrand).**
- **Java also offers good support for graphics which can be easily interfaced with the simulation code. Proper graphics tools, and ways to analyse data interactively, are essential in any simulation project (Alex Nazarenko's contributions were the greatest here)**
- **MONARC simulation and modelling of distributed computing systems provides a realistic description of complex data, and data access patterns and of very large number of jobs running on large scale distributed systems and exchanging very large amount of data (Data Model developed by Krzysztof Sliwa and Iosif Legrand)**

Baseline Model for Daily Activities

Physics Group Analysis	**200-400 jobs/day**
Physics Group Selection	**20-40 jobs/day**
Reconstruction ESD	**2 times/year**
Redefinition of AOD+TAG	**once/month**
Replication (FTP)	**after Reconstruction**
Monte-Carlo	

Event processing rate: 1, 000, 000, 000 events/day

Regional Center Model



VALIDATION MEASUREMENTS

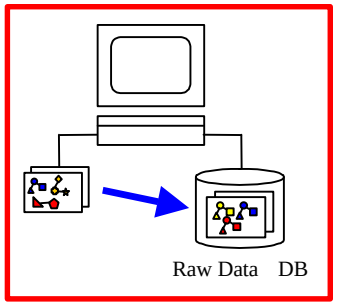
Multiple jobs reading concurrently objects from a data base.

➪Object Size = 10.8 MB

Local DB access

"**atlobj02**"-local

2 CPUs x 300MHz

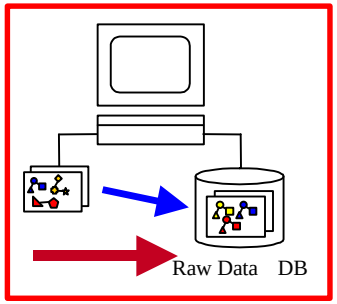


DB on local disk

13.05 SI95/CPU

"**monarc01**"-local

4 CPUs x 400MHz

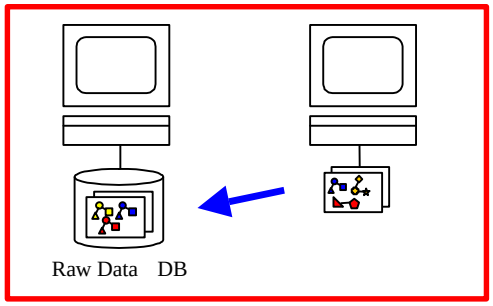


DB on local disk

17.4 SI95/CPU

DB access via AMS

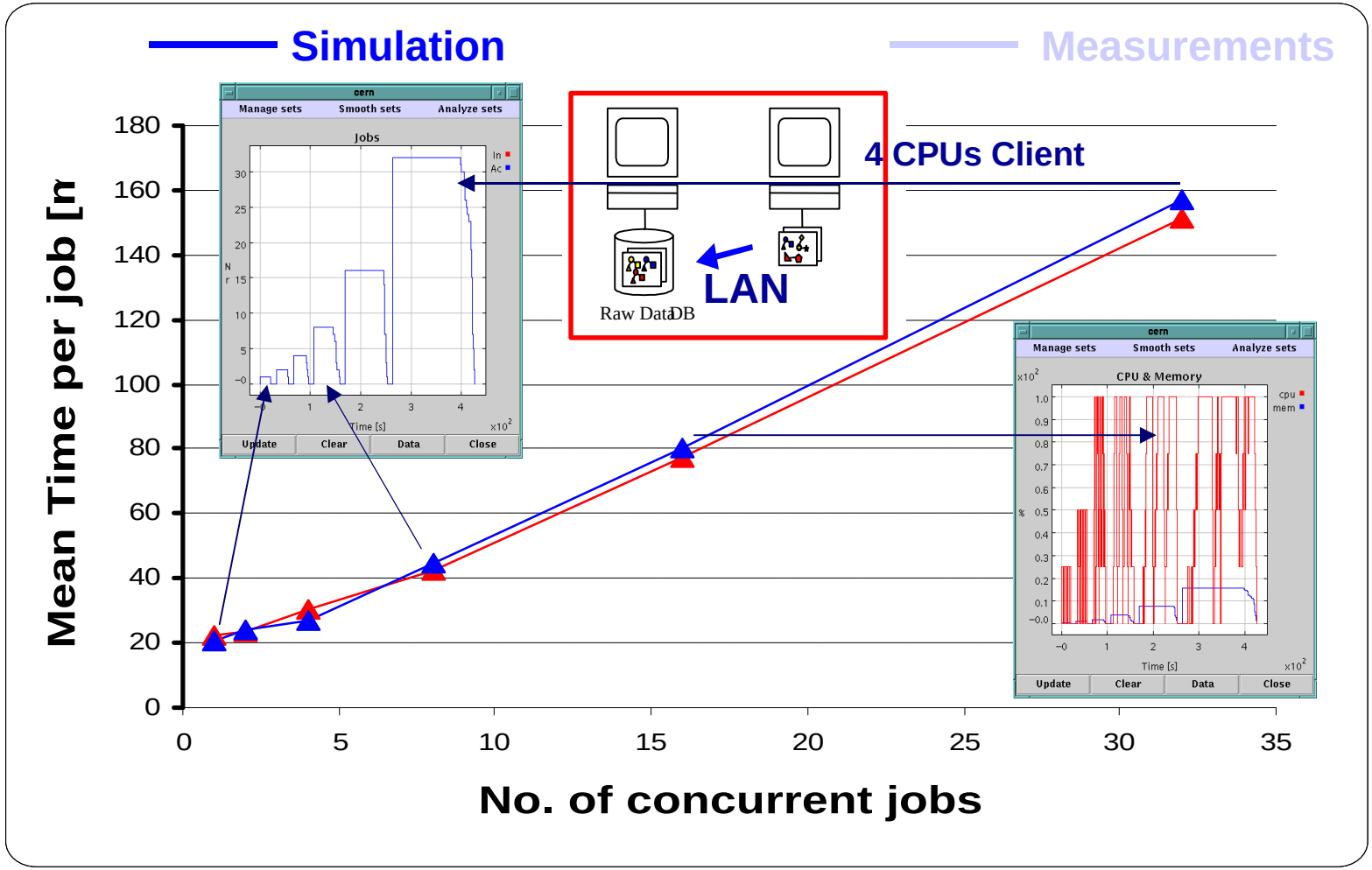
server : "atlobj02"
client : "monarc01"



DB on AMS Server

monarc01 is a 4 CPUs SMP machine
atlobj02 is a 2 CPUs SMP machine

Validation Measurements I
The AMS Data Access Case

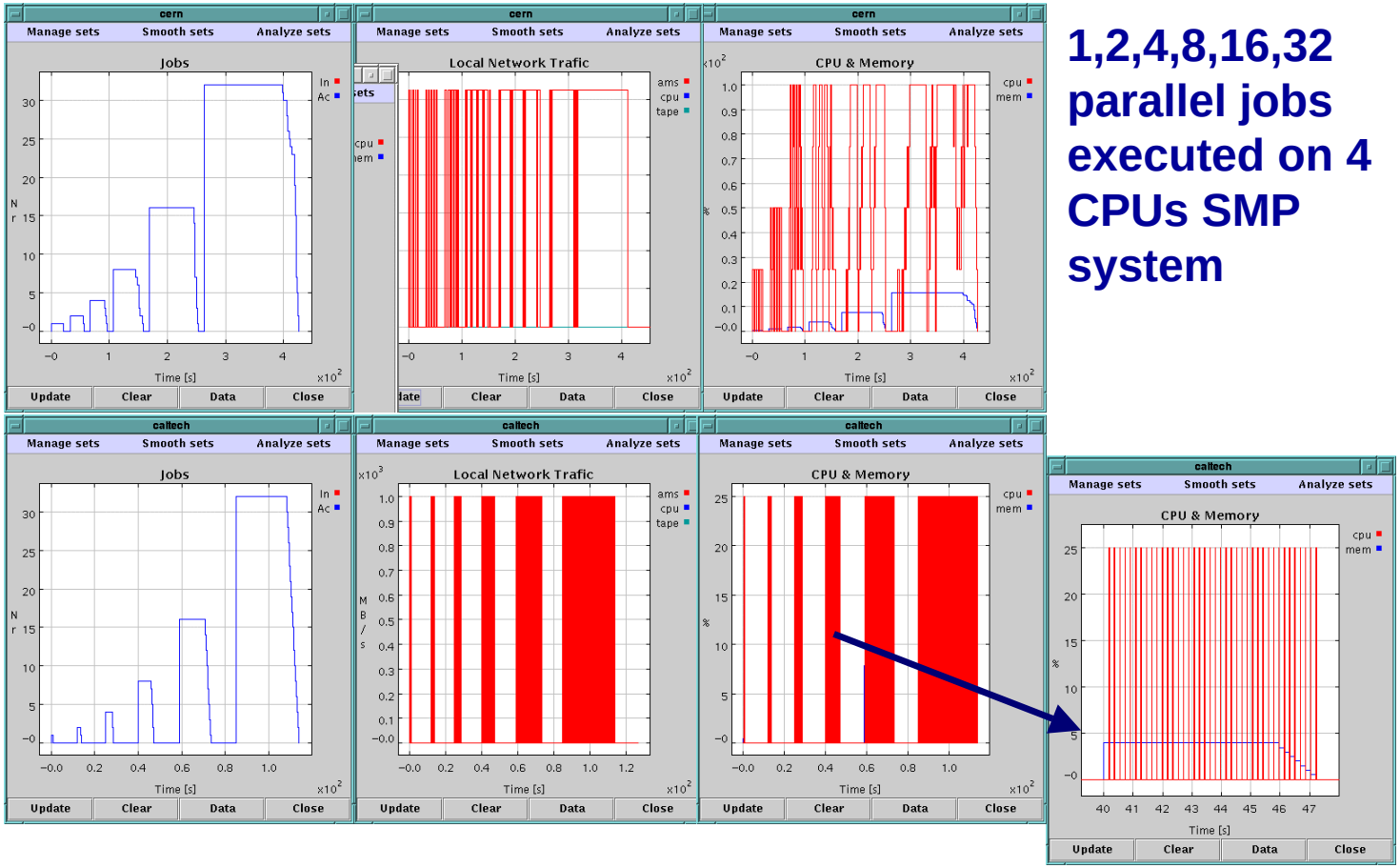


Validation Measurements I
Simulation Results

Simulation results for AMS & Local Data Access

DB access via AMS

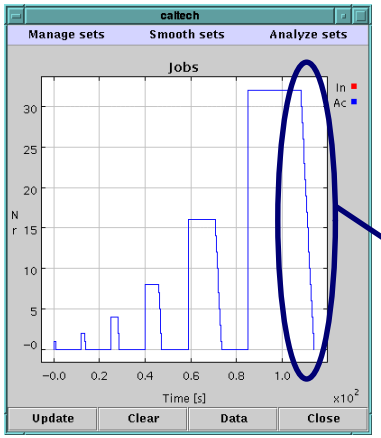
Local DB access



1,2,4,8,16,32 parallel jobs executed on 4 CPUs SMP system

Validation Measurements I

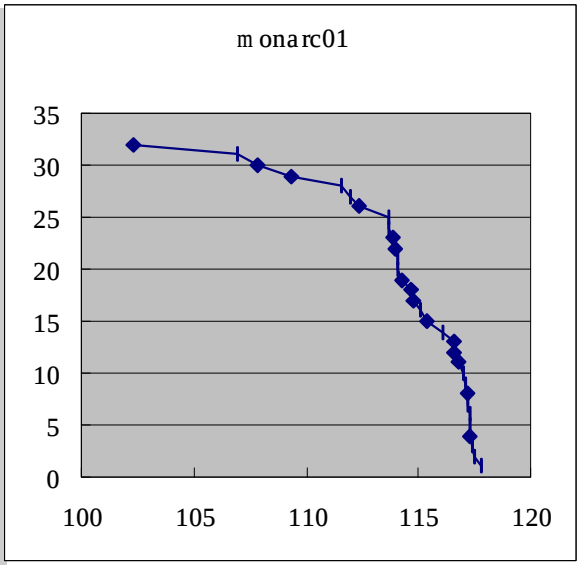
The Distribution of the jobs processing time



Simulation mean 109.5

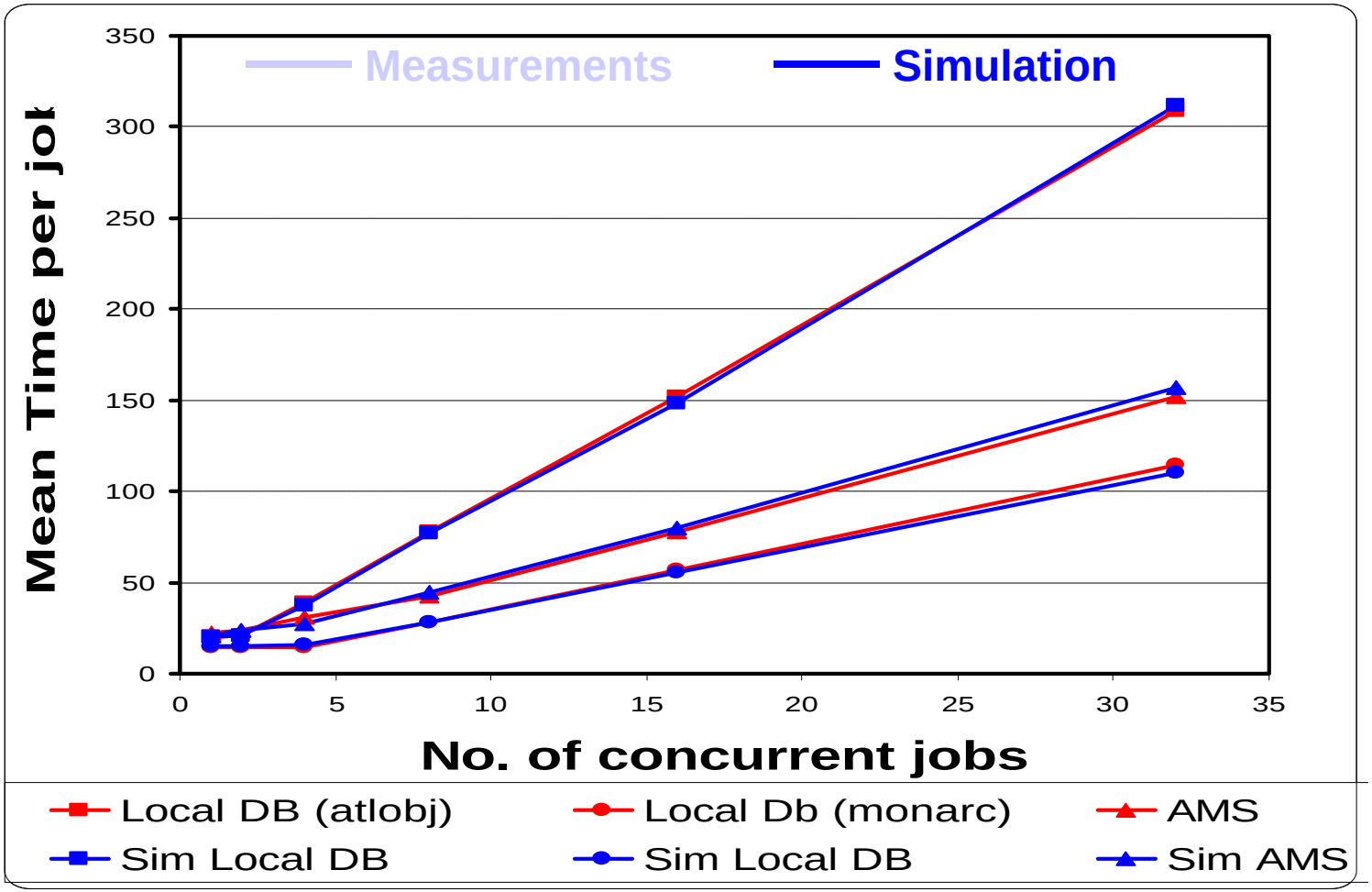
Measurement mean 114.3

Local DB access 32 jobs



Validation Measurements I

Measurements & Simulation

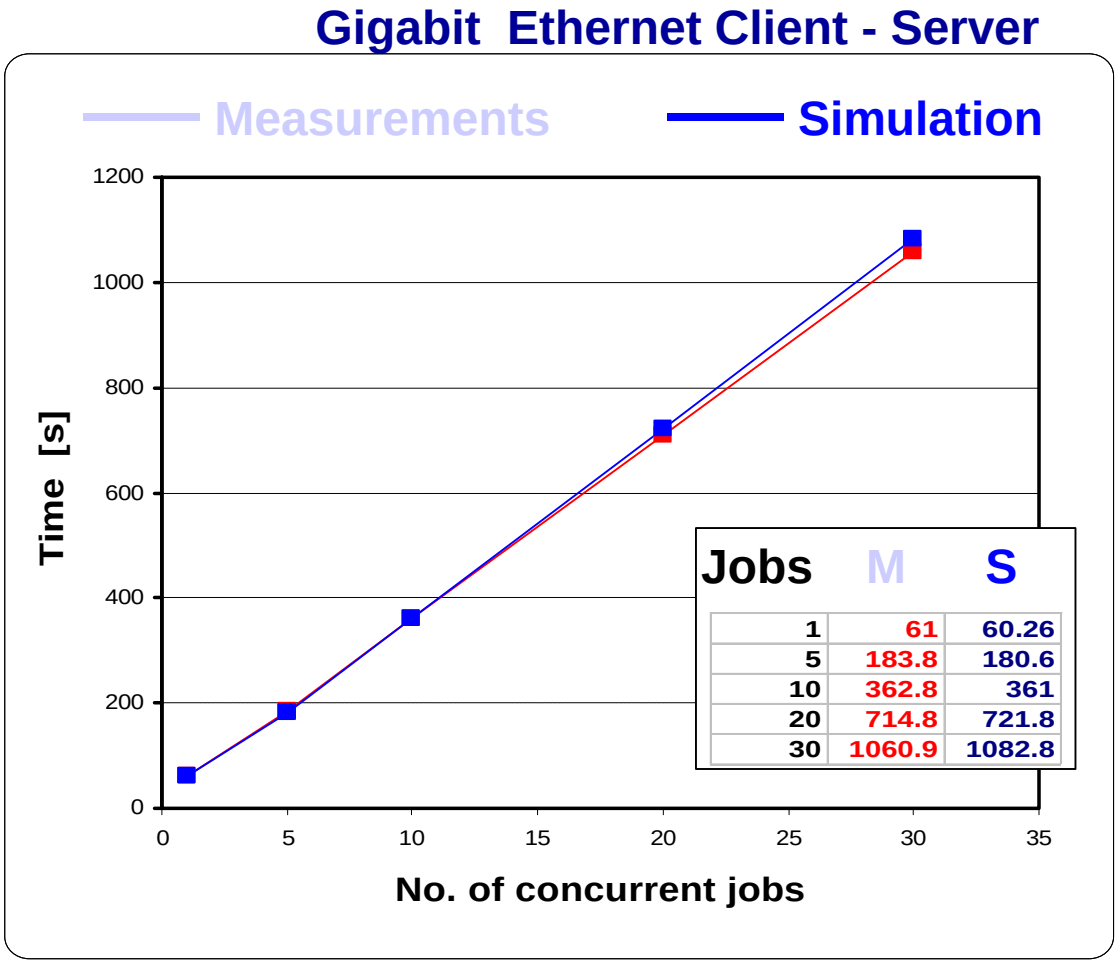


Validation Measurements II

Test 1 **CNAF**	**sunlab1** **Sun Ultra5, 333 MHz**	**1000BaseT**	**gsun** **Sun Ultra5, 333 MHz**
Test 2 **PADOVA**	**cmssun4** **Sun Ultra10, 333 MHz**	**10BaseT**	**vlsi06** **Sun Sparc20, 125 MHz**
Test 3 **CNAF-CERN**	**sunlab1** **Sun Ultra15, 333 MHz**	**2 Mbps**	**monarc01** **Sun Enterprise 4X450 MHz**

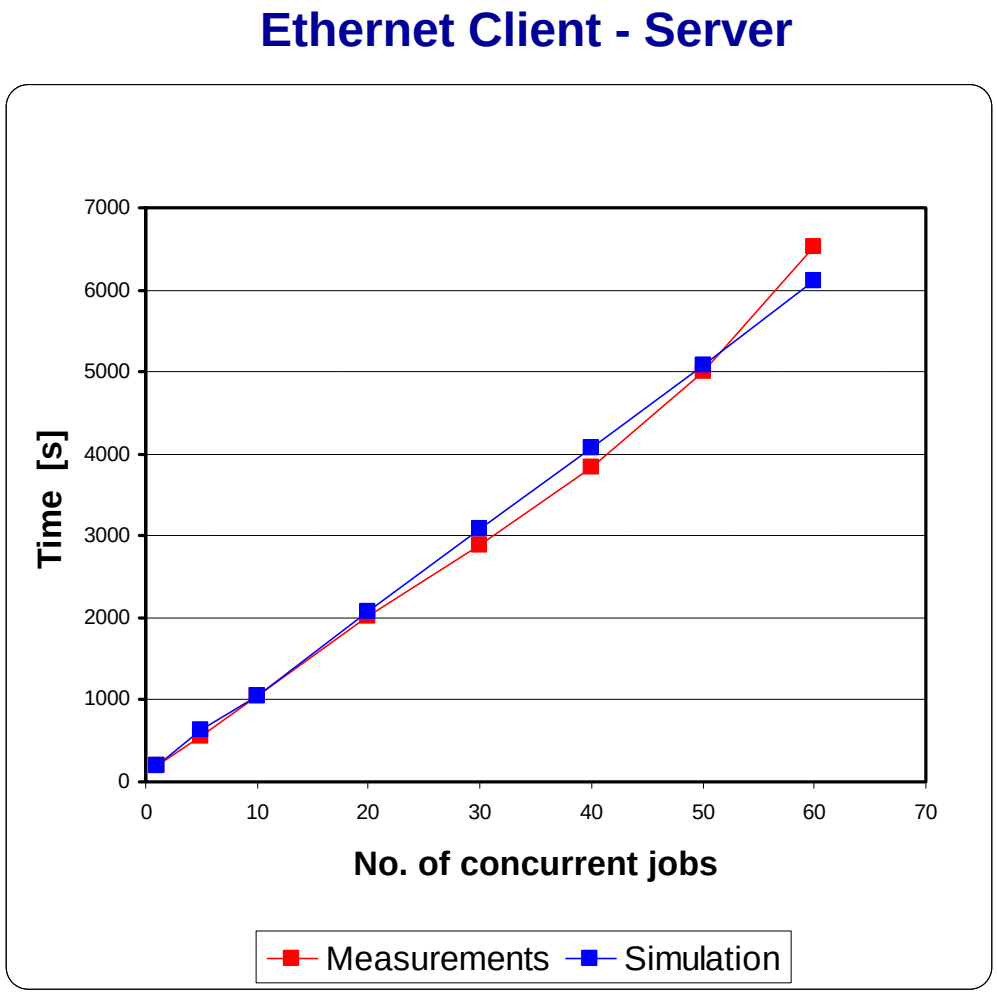
Validation Measurements II

Test 1



Validation Measurements II

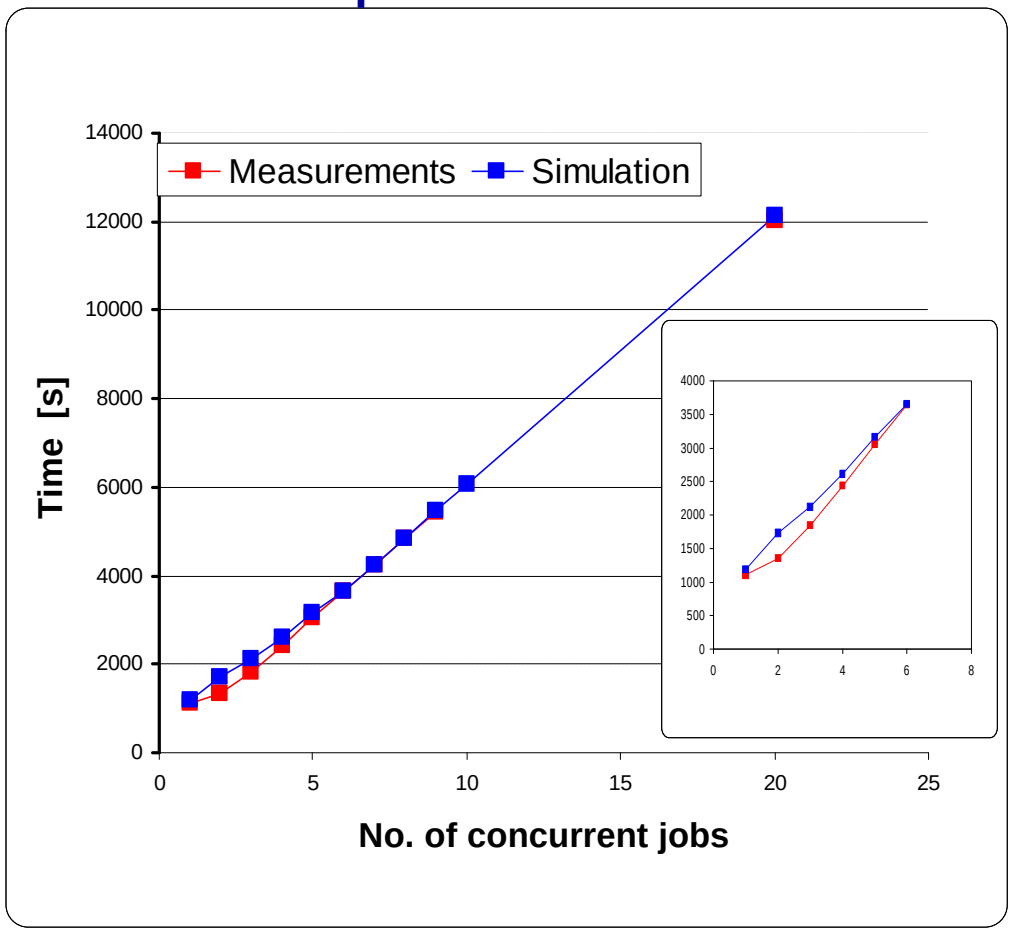
Test 2



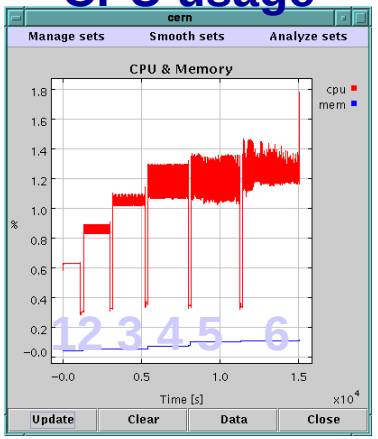
Validation Measurements II

Test 3

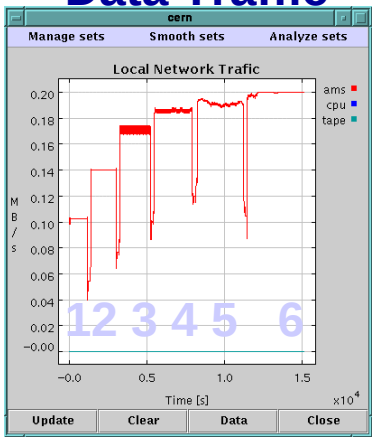
2Mbps WAN Client - Server



CPU usage

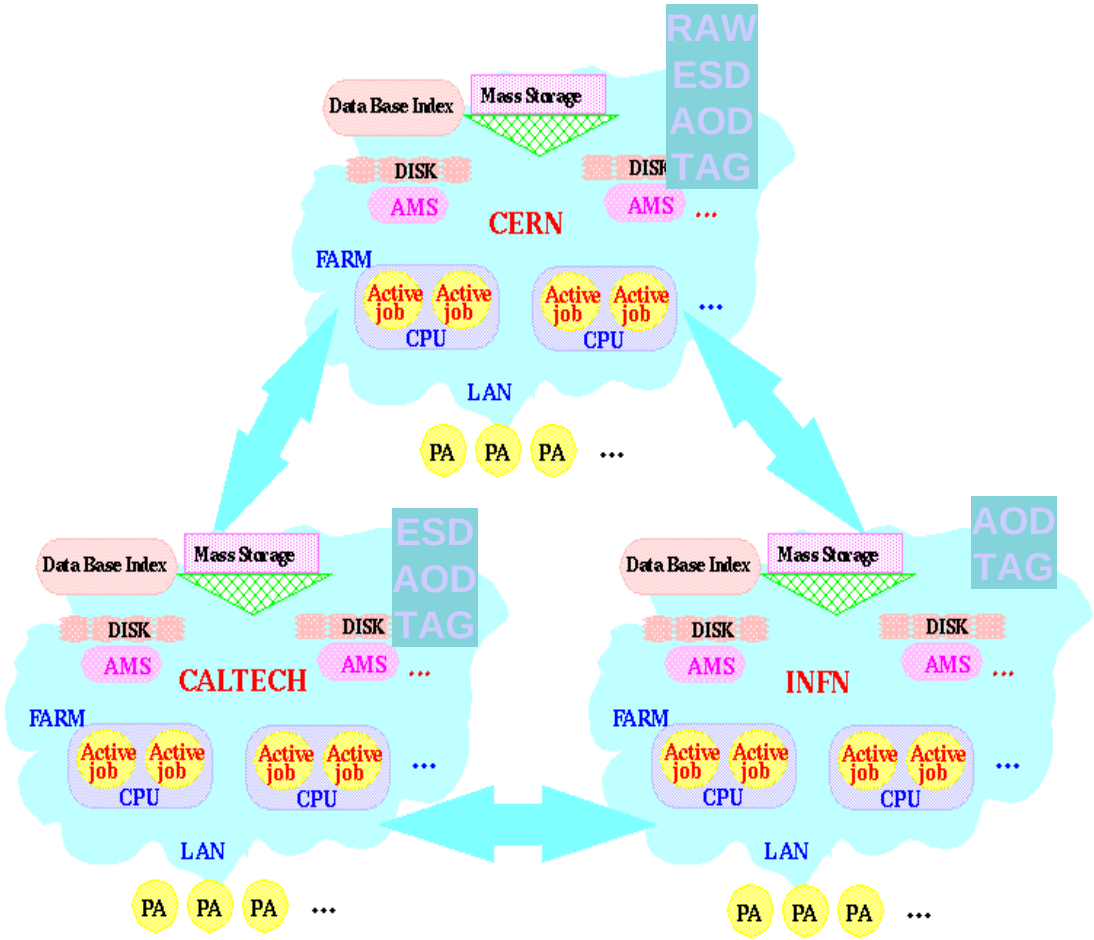


Data Traffic



Physics Analysis Example

- **Similar data processing jobs are performed in six RCs**
- **CERN has RAW, ESD, AOD, TAG**
- **CALTECH, INFN, Tufts, KEK has ESD, AOD, TAG**
- **"Caltech2" has AOD, TAG**



Physics Group Selection

Each group reads 100% TAG events and follows:

~10% to AOD **~1% to ESD** **~0.01% to RAW**

Number of Groups	Follow AOD	Jobs /group
L Groups (L~10)	**p%** of total TAG (~1%)	**1-2**
M Groups (M~5)	**q%** of total TAG (~5%)	**1-2**
N Groups (N~5)	**r%** of total TAG (~10%)	**1-2**

~20 Jobs/Day in total evenly spread among participating RCs

2005-2006 Estimate of Parameters

Parameter	2005	2006
Total CPU	350,000 SI95	520,000 SI95
CPU Unit	400 SI95/box (100 SI95/cpu)	400 SI95/box (100 SI95/CPU)
CPU I/O	40 MBps/box (0.1 MBps/SI95)	40 MBps/box (0.1 MBps/SI95)
AMS I/O for Discs		188 MBps/server
throughput for Tape Storage		2000 MBps
Disk Space	340 TB	540 TB
Tape Space	1 PB	3 PB
LAN	31 GBps	46 GBps

(Les Robertson's estimate of July 99)

Problem Setting: Analysis and Selection

	RAW	ESD	AOD	TAG
Database	1,000,000,000 CERN	1,000,000,000 Tier 1:Locally Tier2:@Tier1	1,000,000,000 Locally @ RC	1,000,000,000 Locally @ RC
Physics Group Analysis 20 groups * 10 jobs	0.01%	1%	Follow 100% of the group set	Group set: 1% of total TAG
Physics Group Selection 20 groups * 1job	0.01%	1%	10%	100%
CPU (SI95)	250	25	2.5	.25

Totally 220 Independent Jobs: 200 Physics Group Analysis and 20 Group Selection

Problem Setting: Reconstruction and FTP

	Size	ESD	AOD	TAG
FTP	1 DB/Job	Tier 1 Centers	Tier 1 & 2	Tier 1 & 2
Full Reconstruction	6,000,000 events/day	yes	yes	yes
Monthly Reconstruction	100,000,000 events/day	no	yes	yes
CPU (SI95)		250	25	2.5

Participating Regional Centers

5 Tier 1 Regional Centers and one Tier 2 center

RC Name	Data	WAN Connection
CERN (Tier1)	RAW, ESD, AOD, TAG	All RCs
INFN (Tier1)	ESD, AOD, TAG	All Tier 1
KEK (Tier1)	ESD, AOD, TAG	All Tier 1
TUFTS (Tier1)	ESD, AOD, TAG	All Tier 1
CALTECH (Tier1)	ESD, AOD, TAG	All Tier 1 & Caltech-2
CALTECH-2 (Tier2)	AOD, TAG	CERN (RAW) & CALTECH (ESD)

200 Analysis and 20 Selection Jobs are evenly spread among Tier1 RCs

AMS load distribution

One RC (CERN) configured to run 200 concurrent Physics Group Analysis Jobs and 20 Selection jobs a day

Participating RC	Data	Jobs
CERN (Tier1)	**RAW, ESD, AOD, TAG**	**200** Physics Group Analysis **20** Physics Group Selection x40

Model1 **(optimized AMS distribution)** Model2

1 RC Vs 5 RC: Group Selection on all data

Model1

One RC (CERN) minimally configured to run 20 concurrent Physics Group Selection Jobs a day

Participating RC	Data	Jobs
CERN (Tier1)	**RAW, ESD, AOD, TAG**	**20** Physics Group Selection **x10**

Model2

Five Tier 1 Centers minimally configured to perform the same task

Participating RC	Data	Jobs
CERN	**RAW, ESD, AOD, TAG**	**4** Physics Group Analysis **x10**
INFN	**AOD, TAG**	**4** Physics Group Analysis **x10**
KEK	**AOD, TAG**	**4** Physics Group Analysis **x10**
TUFTS	**AOD, TAG**	**4** Physics Group Analysis **x10**
CALTECH	**AOD, TAG**	**4** Physics Group Analysis **x10**

Conclusion: **Current configuration provides a possibility to redistribute resources without much increase in the cost; further optimization is needed to increase the efficiency: *1/(Time*Cost)***

1 RC vs 6 RC: Reconstruction+Analysis+Selection

Model1

One RC (CERN) minimally configured to run all the Jobs a day

Participating RC	Data	Jobs
CERN (Tier1)	**RAW, ESD, AOD, TAG**	**20** Physics Group Selection **x10** **200** Physics Group Analysis Full Reconstruction and FTP

Model2

Five Tier 1 Centers optimized to perform the same task with 30 MBps WAN

Participating RC	Data	Jobs
CERN	**RAW, ESD, AOD, TAG**	**4** Physics Group Selection**x10** **40** Physics Group Analysis Full Reconstruction and FTP
INFN	**ESD, AOD, TAG**	**4** Physics Group Selection**x10** **40** Physics Group Analysis
KEK	**ESD, AOD, TAG**	**4** Physics Group Selection**x10** **40** Physics Group Analysis
TUFTS	**ESD, AOD, TAG**	**4** Physics Group Selection**x10** **40** Physics Group Analysis
CALTECH	**ESD, AOD, TAG**	**4** Physics Group Selection**x10** **40** Physics Group Analysis
CALTECH-2 (Tier2)	**AOD, TAG**	**20** Physics Group Analysis

Conclusion: **Current configuration provides a possibility to optimize the CPU power and reduce the co**
further optimization is possible to reduce WAN bandwidth to 30 MBps

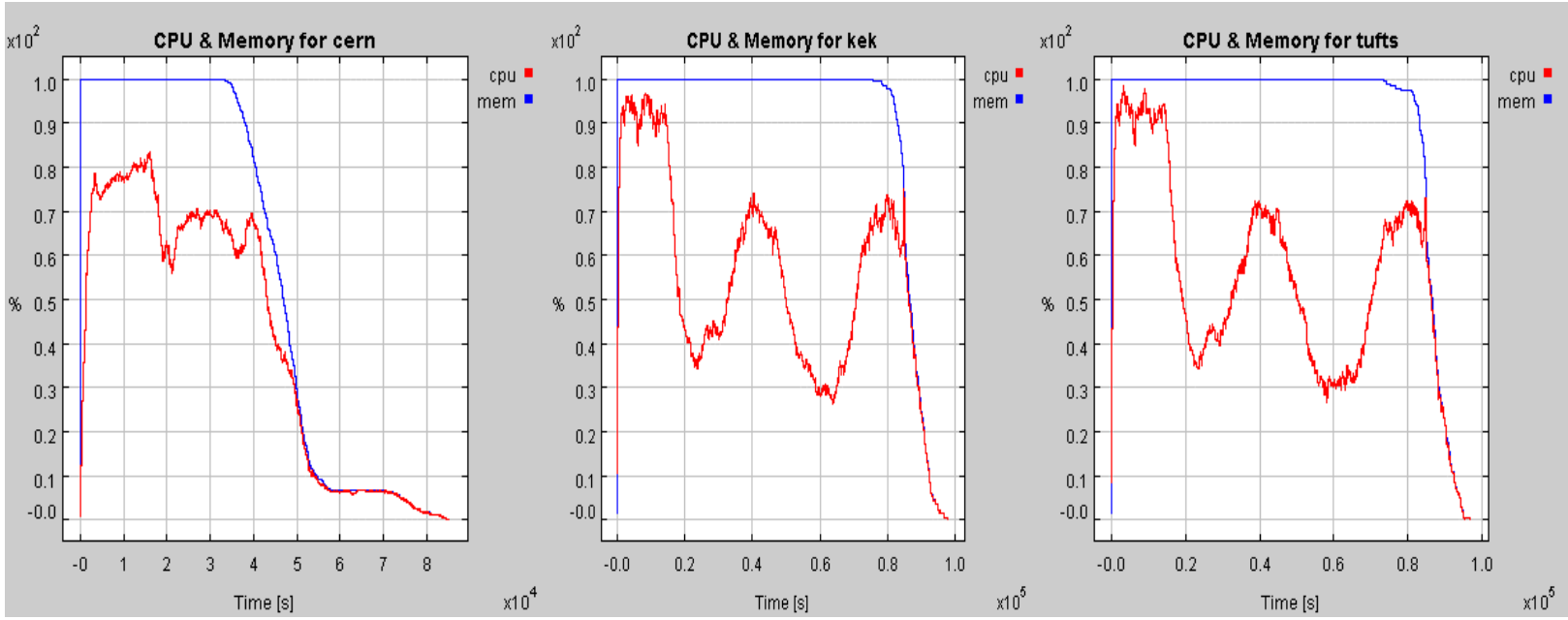
6 RC: Two types of Reconstruction+Analysis+Selection

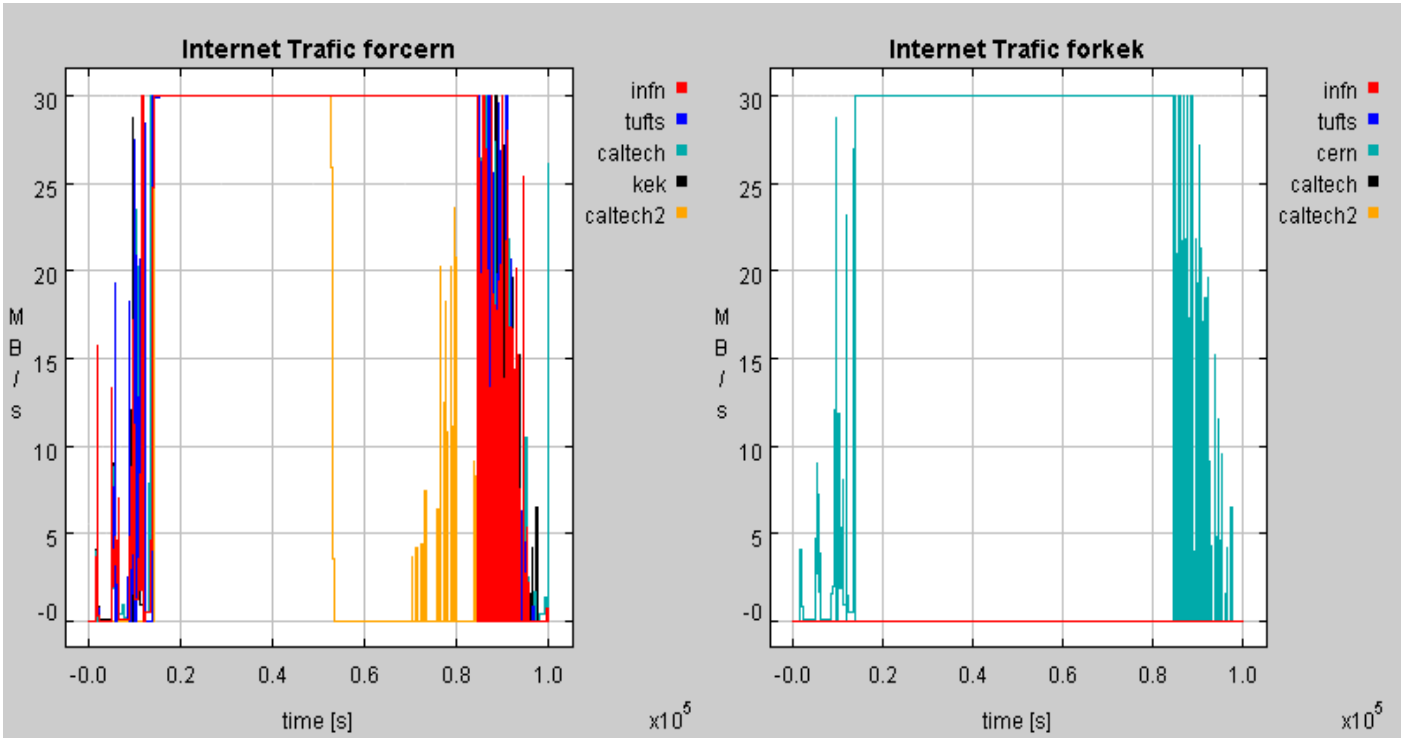
Five Tier 1 and one Tier 2 Centers optimized to perform the complete set with 30 MBps WAN and optimized LAN

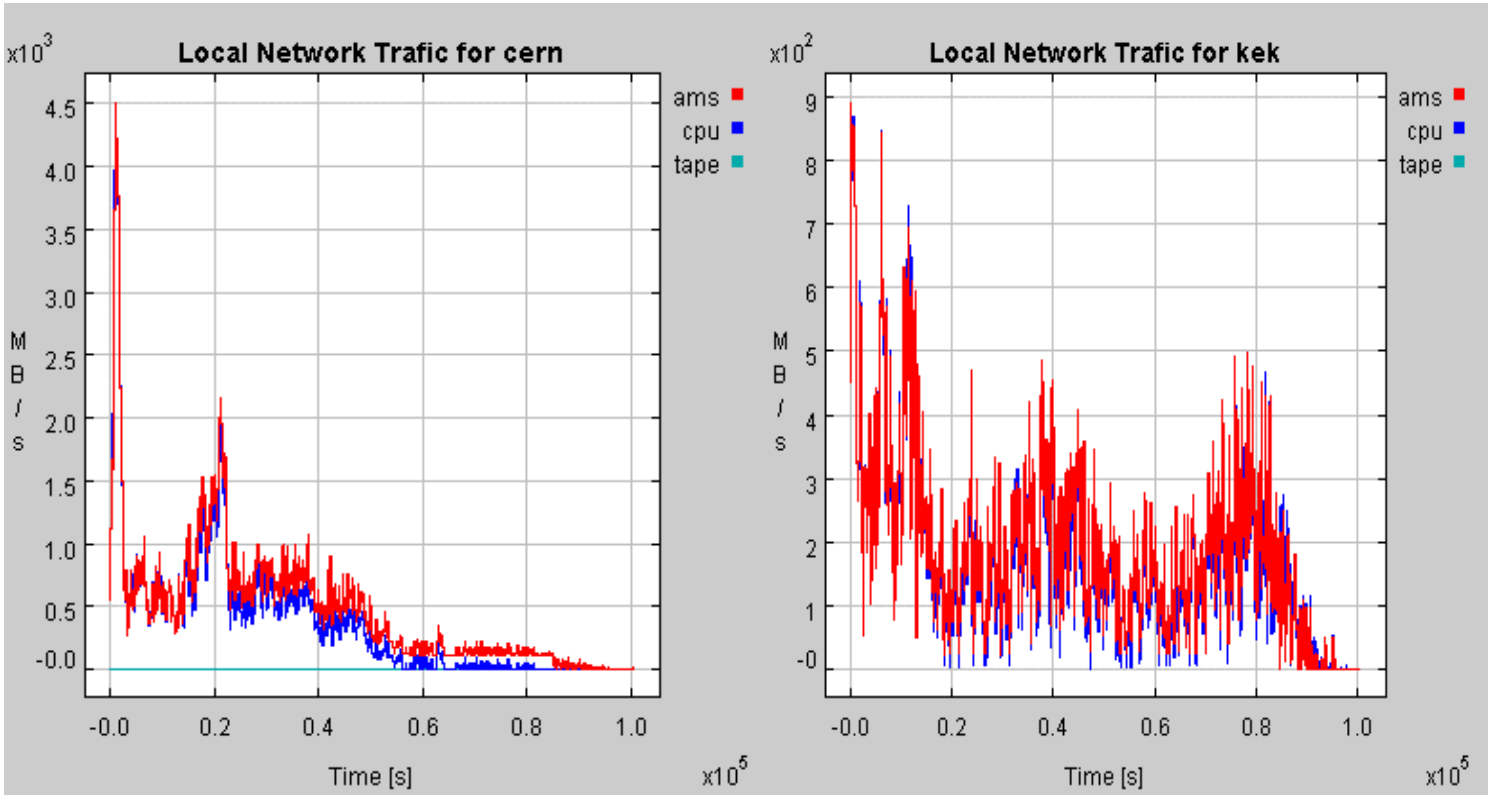
Participating RC	Data	Jobs
CERN	**RAW, ESD, AOD, TAG**	**4** Physics Group Selection**x10** **40** Physics Group Analysis Full Reconstruction and FTP
INFN	**ESD, AOD, TAG**	Monthly Reconstruction and FTP (10days)
KEK	**ESD, AOD, TAG**	**4** Physics Group Selection**x10** **40** Physics Group Analysis
TUFTS	**ESD, AOD, TAG**	**4** Physics Group Selection**x10** **40** Physics Group Analysis
CALTECH	**ESD, AOD, TAG**	**4** Physics Group Selection**x10** **40** Physics Group Analysis
CALTECH-2	**AOD, TAG**	**4** Physics Group Selection**x10** **40** Physics Group Analysis **20** Physics Group Analysis

Model1 **(fixed values)** Model2 **(randomized data processing times and sizes)**

Conclusion: **Current configuration provides a possibility to run daily the complete set of jobs at 6 centers with the WAN bandwidth 30 MBps and the network parameters not exceeding the estimate of 2005**


CPU & Memory for cern
CPU & Memory for kek
CPU & Memory for tufts
cpu
mem
%
Time [s]
x10^2
x10^4
x10^5


Internet Trafic forcern
infn
tufts
caltech
kek
caltech2
30
25
20
15
10
5
-0
M
B
/
s
-0.0
0.2
0.4
0.6
0.8
1.0
time [s]
x10^5
Internet Trafic forkek
infn
tufts
cern
caltech
caltech2
30
25
20
15
10
5
-0
M
B
/
s
-0.0
0.2
0.4
0.6
0.8
1.0
time [s]
x10^5


Local Network Trafic for cern
x10^3
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
-0.0
MB/s
-0.0
0.2
0.4
0.6
0.8
1.0
Time [s]
x10^5
ams
cpu
tape
Local Network Trafic for kek
x10^2
9
8
7
6
5
4
3
2
1
-0
MB/s
-0.0
0.2
0.4
0.6
0.8
1.0
Time [s]
x10^5
ams
cpu
tape

Future Work:

- **Replication job: partial dataset replication from CERN to Regional Centers**
- **Flexible data access: data exchange between Regional Centers without getting data directly from CERN (depending on load, availability,...)**
- **Imposing coherence on the concurrent jobs: if Reconstruction and/or Replication is taking place, Analysis/Selection jobs should be able to monitor new data availability if requested**
- **Improving Cost function: adding cost of WAN, adding other hidden costs currently not accounted for**
- **Optimization with respect to the parameter *Time*Cost* for a given task run on different architectures**